Near-separable Non-negative Matrix Factorization with ℓ_1- and Bregman Loss Functions

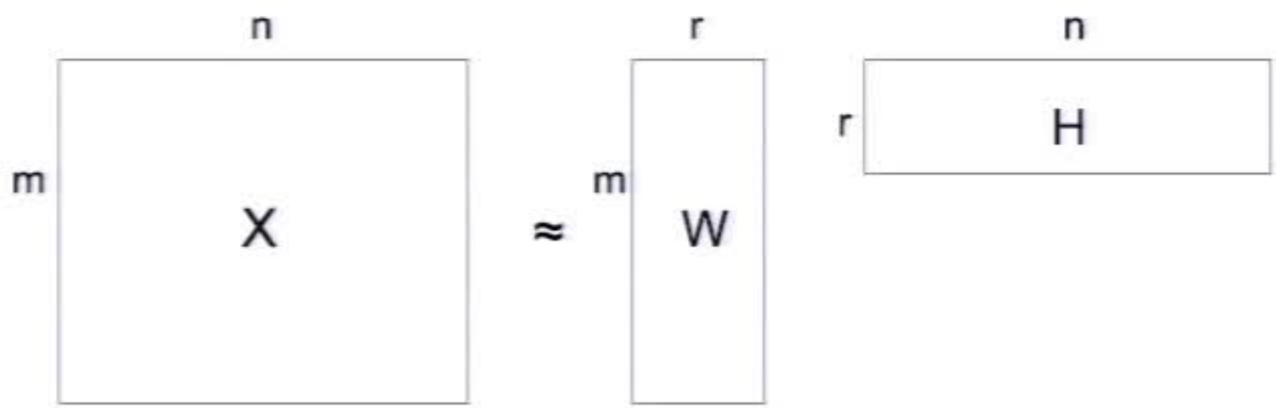
Abhishek Kumar[a] and Vikas Sindhwani[v]

Presented by Kush R. Varshney[k]

[a,k] IBM TJ Watson Research Center, Yorktown Heights, NY
[v] Google Research, NY

Nonnegative Matrix Factorization (NMF)

- Given a matrix $X \geq 0$, express it as a product WH (W, $H \geq 0$)



  – Minimize some distance measure between X and WH
  – The inner dimension r is less than both m and n.
- Nonnegative rank: Smallest inner dimension r s.t. $X = WH$

  $\text{Rank}(X) \leq \text{Rank}_+(\text{X}) \leq \min(m, n)$

- Non-unique: The factorization can be non-unique (even after ignoring permutation and scaling).

Motivation: Decomposition into Parts

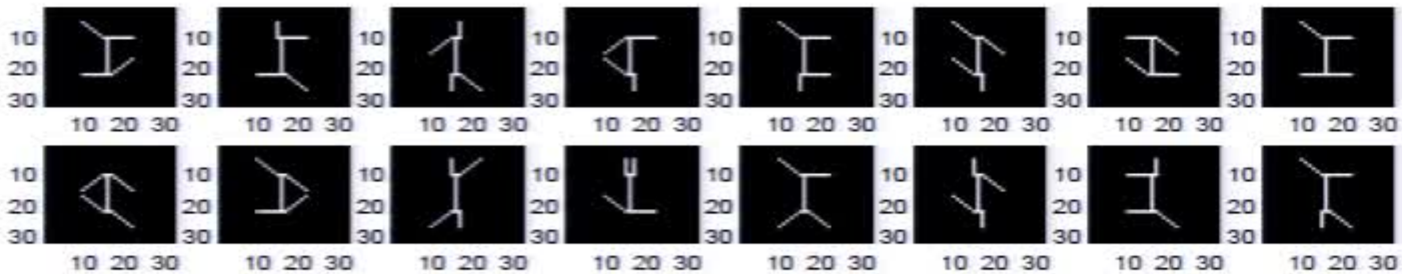
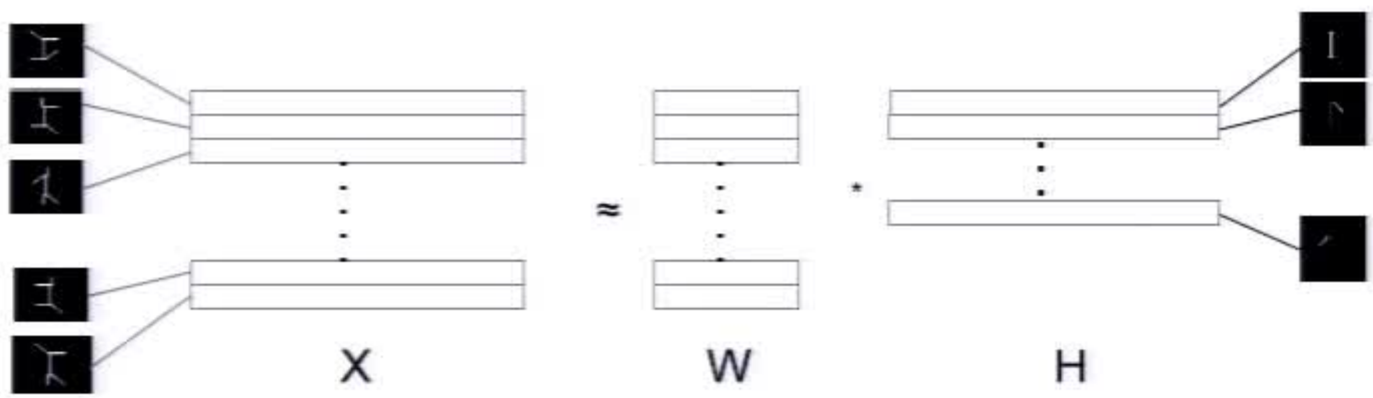


Figure : Sample images from Swimmer database (256 images, each of size 32×32)



H: basis / *topics* / dictionary, **W**: reconstruction coefficients

- Nonnegative **W** implies additive combination of topics

Motivation: Decomposition into Parts

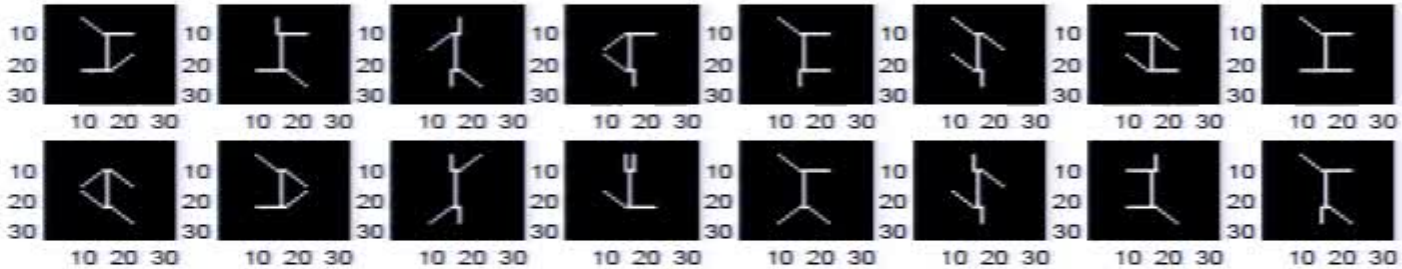
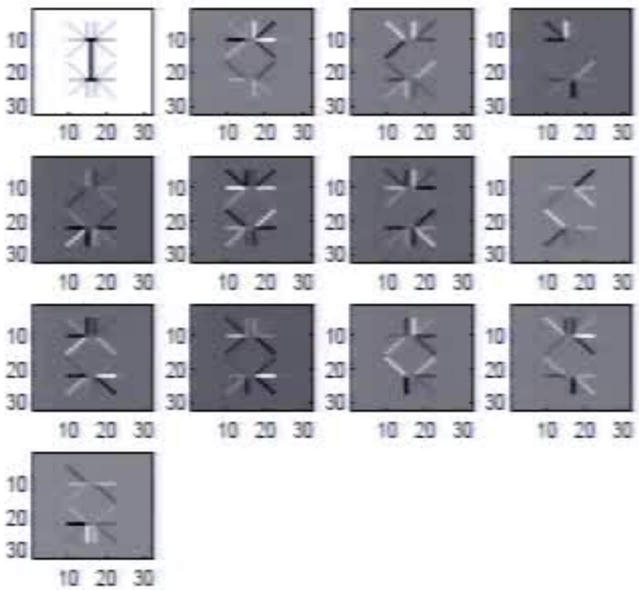


Figure : Sample images from Swimmer database (256 images, each of size 32×32)



(a) Basis obtained by SVD: $\min_{W,H} \|X - WH\|_F$

Motivation: Decomposition into Parts

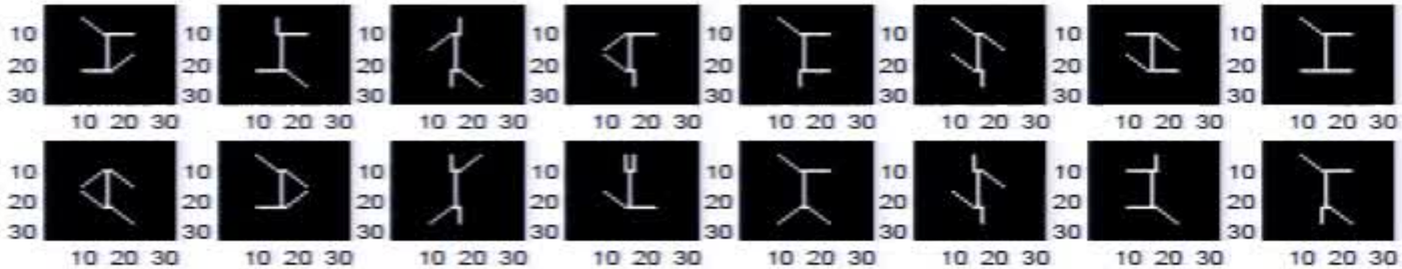
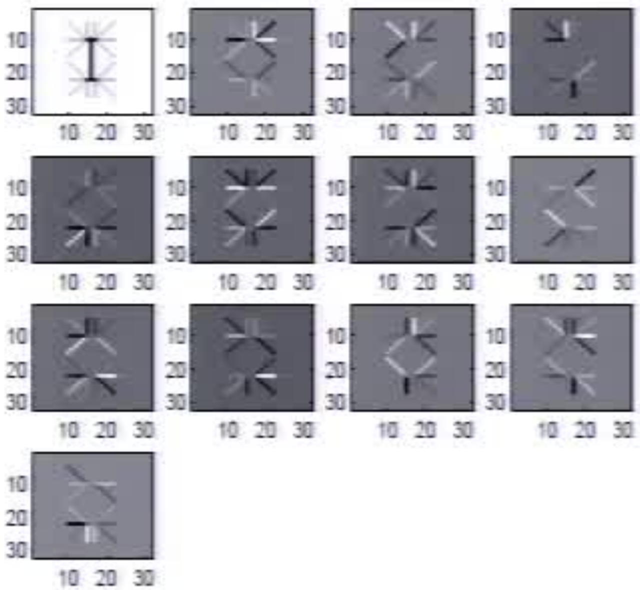
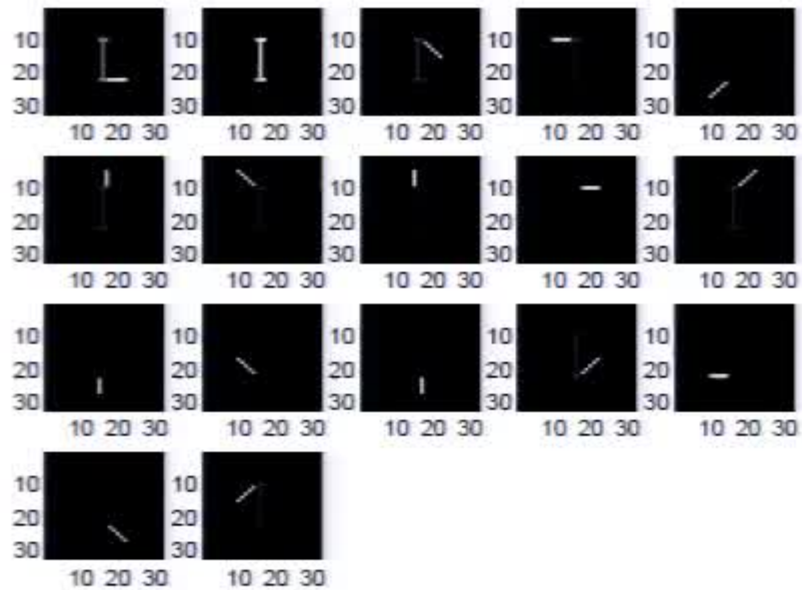


Figure : Sample images from Swimmer database (256 images, each of size 32×32)



(a) Basis obtained by SVD: $\min_{W,H} \|X - WH\|_F$



(b) “Topics” obtained by NMF: $\min_{W,H \geq 0} \|X - WH\|_F$

NP-hardness and Existing local search based approaches

- NMF problem is NP-hard (Vavasis, 2009) hence majority of work in the area has been on local-search methods.

$$\min_{W,H} L(X, WH) \qquad \text{s.t. } W \geq 0,\ H \geq 0$$

where $L(\cdot,\cdot)$ is a suitable loss function.

1. Randomly initialize nonnegative W and H
2. Fix one (block) of the variables and optimize for the others.
3. Cycle over the blocks (block coordinate descent).

- Guaranteed to converge to a stationary point, not necessarily to the global optimum.
- Several flavors exist to make the optimization fast.

Separability assumption

- **Assumption:** X is r-separable if identity matrix (I) is hidden in H

$$X = W_{m\times r} \underbrace{\left[I_{r\times r} \; H'_{r\times(n-r)} \right] P}_{H} \quad \text{for some permutation matrix } P$$

  – Columns of W appear as it is in X: $W = X(:, A)$

- **Anchors:** those columns of X that appear in W (given by the set A)

  – other columns are conic combination of anchors: $X = X(:, A)H$

- NMF problem reduces to finding the extreme rays of the cone containing columns of X

Separability assumption

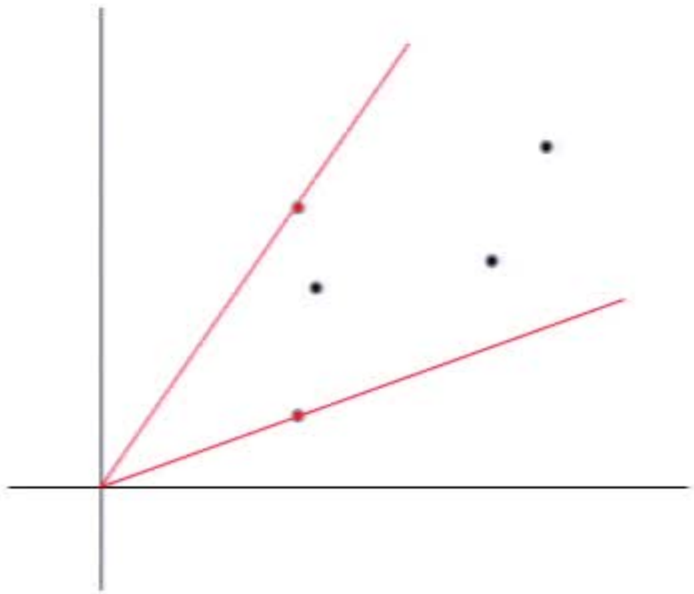


Figure : Red points are anchors

NMF problem: find extreme rays of the pointed polyhedral cone

Near-separable (noisy) problem

- Problem: Given

$$X = \underbrace{W[I_r\ H']P}_{\text{separable structure}} + \underbrace{N}_{\text{noise}}$$

  identify the anchor columns of X.
- There are existing methods that model the noise by Frobenius-norm loss (Kumar et al, 2013), $\ell_{1,\infty}$-norm loss (Bittorf et al, 2011). Some methods do not model the noise explicitly (Gillis and Vavasis 2014, Arora et al, 2013).
- We propose near-separable NMF with ℓ_1 and general Bregman loss functions to broaden its applicability.
  – ℓ_1- loss models the sparse noise case while Bregman loss functions can model noise from exponential family of distributions.

Separability: a reasonable assumption?

$$X \approx W_{m\times r} \underbrace{\left[I_{r\times r} \; H'_{r\times(n-r)} \right] P}_{H} \quad \text{for some permutation matrix } P$$

- Several previous works have shown that separability is a reasonable assumption for Topic modeling (Arora et al, 2013, Kumar et al, 2013) and Hyperspectral unmixing (Gillis and Vavasis, 2014).
- We demonstrate that separable NMF with ℓ_1 loss performs well for video foreground-background separation and is a strong alternative to the popular Robust PCA approach.

Robust low rank approximation

- Robust low-rank approximation: models sparse noise $X = L + S$, S sparse

$$\min_L \|X - L\|_1 + \lambda \operatorname{rank}(L) \quad \text{OR} \quad \min_L \|X - L\|_1, \text{ s.t. } \operatorname{rank}(L) \le r$$

  – Non-convex; believed to be NP-hard

Robust low *nonnegative rank* approximation

- Robust low nonnegative-rank approximation: models sparse noise
  $X = L + S$, S sparse

$$\min_{L \geq 0} \|X - L\|_1, \quad \text{s.t.} \quad \text{nn-rank}(L) \leq r$$

  – also NP-hard

2:06 PM

Making these robust approximations tractable

- Both low-rank and low nonnegative-rank approximations can be used to do foreground-background separation
  - background subtraction in video
  - background topics from document-specific keywords in text
- Robust low-rank approx.: popular approach to make it tractable is convex relaxation

$$\min_L \|X - L\|_1 + \lambda \|L\|_\star$$

  - well studied in the literature under *Robust PCA*
- Robust low-nonnegative rank approx.: We use separability assumption to make it tractable.

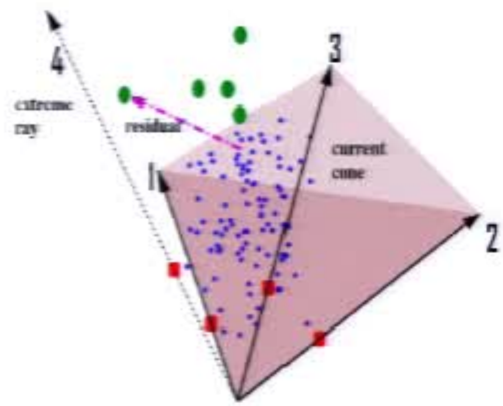
$$\min_{W,H \geq 0} \|X - WH\|_1, \text{ s.t. } H = [I_{r \times r} \; H'_{r \times (n-r)}]P$$

  – How does it compare with convexified Robust PCA on common applications?

Fast nonnegative conic hull

[Kumar et al, ICML 2013]

- Minimize $\|X - WH\|_F^2$ s.t.
  $W \geq 0$, $H = \left[I_{r\times r} \; H'_{r\times(n-r)} \right] P \geq 0$ for some permutation matrix P
- Based on characterization of extreme points/rays from Clarkson, 1994 and Dula et al., 1998

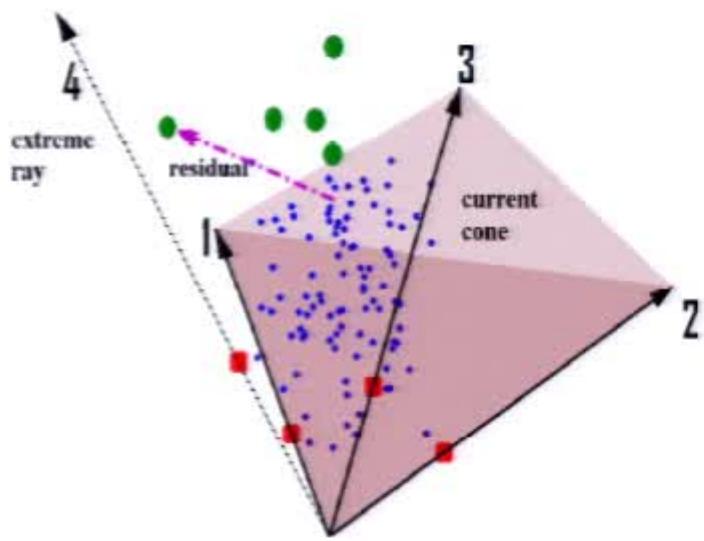


To expand the current cone:

- Step 1: Project external points to the cone and find normals to the faces
- Step 2: Pick a face and rotate until it hits the "last" point
- Step 3: Add this point as a new extreme ray

Nonnegative Conical Hull with ℓ_1 loss

- We extend the Conical hull approach of (Kumar et al, 2013) to ℓ_1 and Bregman loss functions.



To expand the current cone:

- Step 1: Project external points to the current cone
  - ℓ_1 projections are not normal to the faces
- Step 2: Pick a new extreme ray and expand the cone

Nonnegative Conical Hull with ℓ_1 loss

Robust XRAY algorithm:

0. Start: $A = []$, $D \leftarrow X$

1. Selection step: select a column using the following criteria

$$j^* = \arg\max_j \left(\frac{D_i^T X_j}{p^T X_j} \right), \qquad A \leftarrow A \cup j^*$$

2. Project on the cone: Solve multivariate nonnegative least absolute deviation

$$H^* = \arg\min_{H \geq 0} \|X - X(:, A)H\|_1 \qquad \text{(ADMM)}$$

$$D_{ij} \leftarrow \begin{cases} \text{sign}(X - X(:, A)H^*)_{ij}, & \text{if } (X - X(:, A)H^*)_{ij} \neq 0 \\ 0, & \text{if } (X - X(:, A)H^*)_{ij} = 0 \end{cases}$$

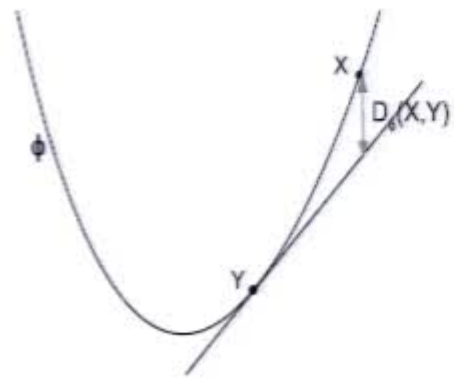
3. Goto step 1 until $|A| = r$

- Provably solves the separable problem.
- Several possibilities in selecting the exterior point.

Nonnegative Conical Hull with Bregman Divergences

- A strictly convex function: $\phi : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$
  Continuous derivative: $\psi : \mathbb{R}^{m\times n} \mapsto \mathbb{R}^{m\times n}$
  Bregman divergence is defined as

$$D_\phi(X, Y) = \phi(X) - \phi(Y) - tr(\psi(Y)^T (X - Y)) \geq 0$$



- Selection criteria can be modified to recover anchor columns with Bregman divergences.

$$\min_{W,H\geq 0} D_\phi(X, WH), \text{ s.t. } H = [I_{r\times r}\ H'_{r\times(n-r)}]P$$

Special case: $\phi(X) = \|X\|_F^2 \Longrightarrow D_\phi(X, WH) = \|X - WH\|_F^2$

Advantages

- Easy model selection: Previous solutions are contained in the current solution
  - can incrementally add new topics until some criterion is met
- Scalable implementation
  - easy to parallelize
  - compares favorably with Robust PCA in terms of speed

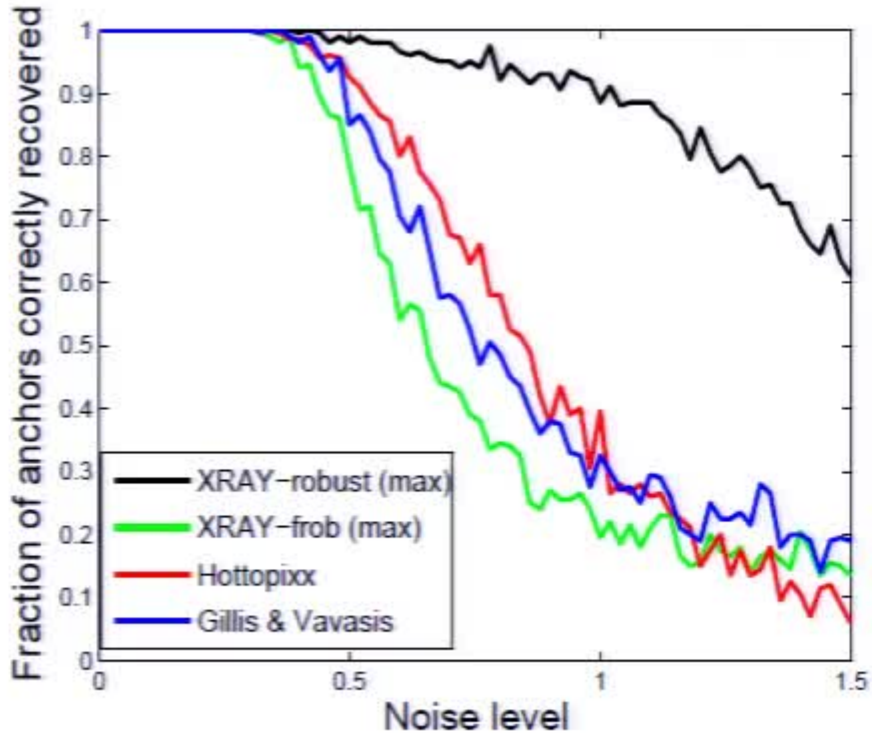
Synthetic data (recovery of anchors): Sparse noise case

Generative model: $X = WH + N$

$W \in \mathbb{R}_+^{200\times 20} \sim$ Uniform between 0 and 1

$H = [I_{20}\ H'] \in \mathbb{R}_+^{20\times 210}$, $H' \sim$ Dirichlet

$N \sim \text{Laplace}(0,\delta)$, $N \leftarrow \max(N, 0)$

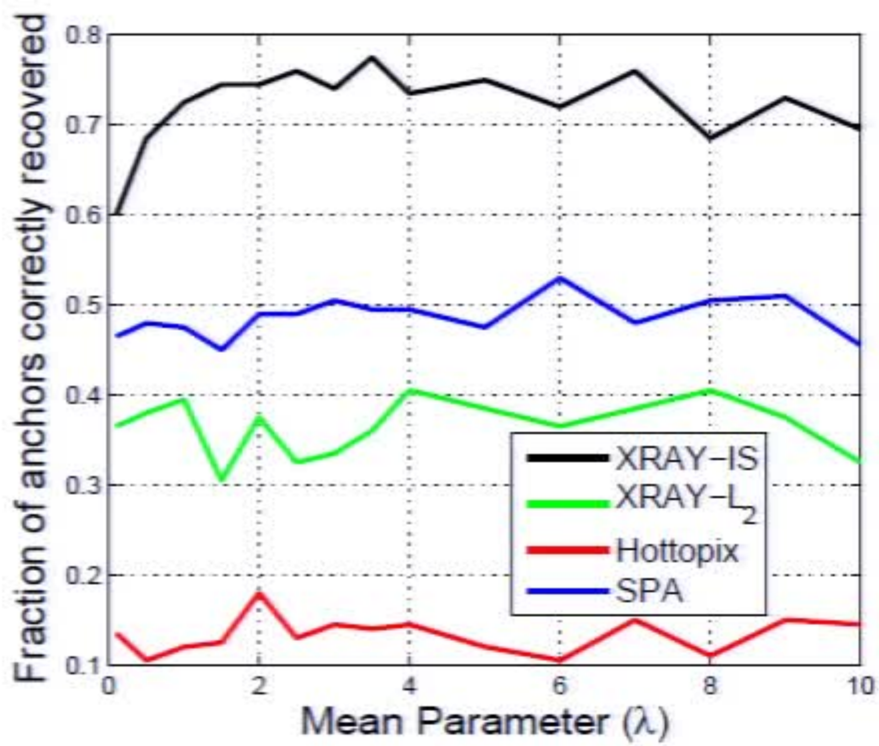


Synthetic data (recovery of anchors): noise from exponential distribution

Generative model: $X_{ij} \sim \exp(\lambda W_{i:} H_{:j})$

$W \in \mathbb{R}_+^{200\times 20} \sim$ Uniform between 0 and 1

$H = [I_{20}\ H'] \in \mathbb{R}_+^{20\times 210}$, $H' \sim$ Dirichlet



Foreground-background separation in Video

$X_{m \times n}$: each row is a video frame

Decompose X into (Low nonnegative-rank) + (Sparse) components using robust separable NMF.

Foreground-background separation in Video

$X_{m \times n}$: each row is a video frame

Decompose X into (Low nonnegative-rank) + (Sparse) components using robust separable NMF.

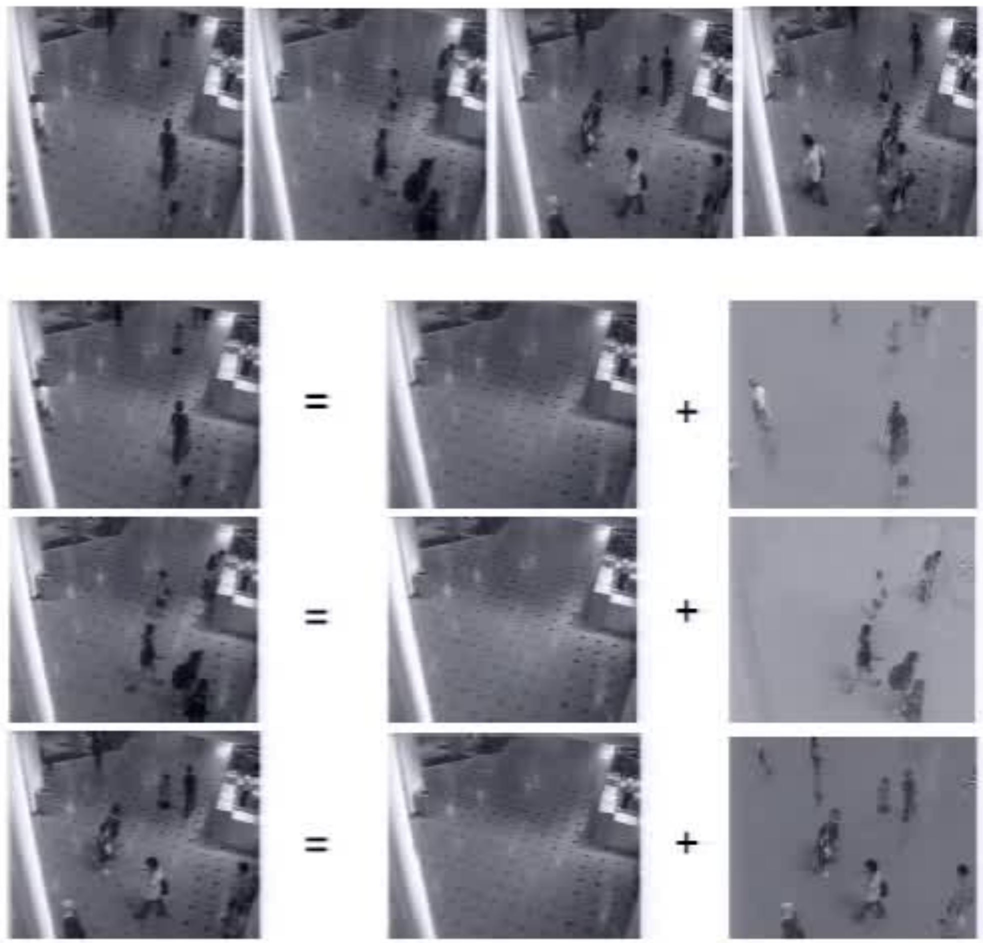
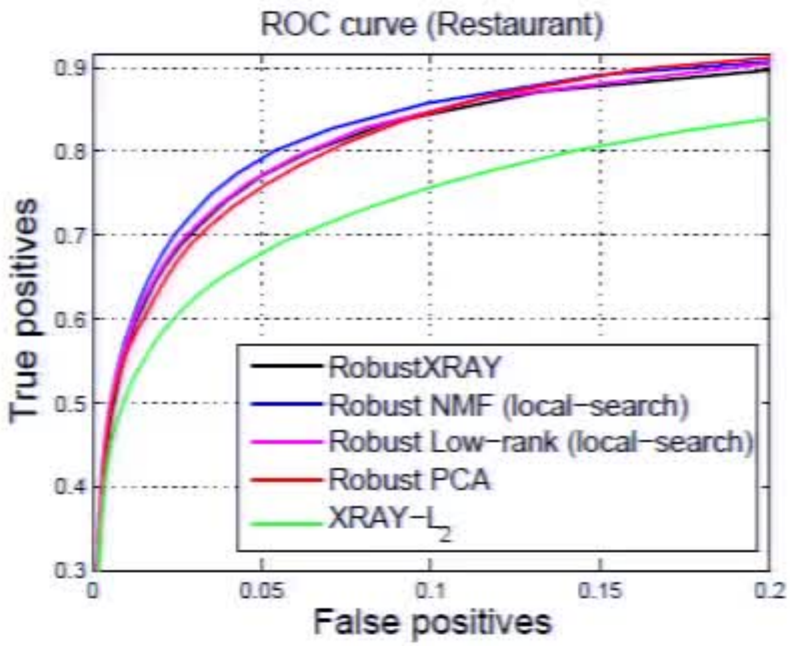


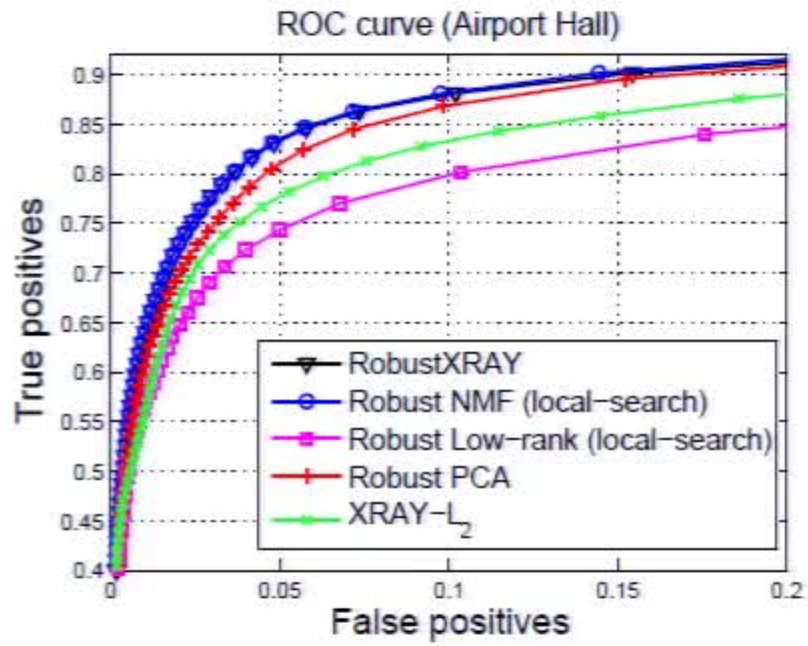
Figure : Background separation using Robust XRAY ($r = 2$)

Foreground-background separation in Video

Robust PCA: $\min_L \|X - L\|_1 + \lambda \|L\|_\star$ (inexact ALM)

Robust Separable NMF: $\min_{W,H\geq 0} \|X - WH\|_1$, s.t. $H = [I_{r\times r}\ H'_{r\times(n-r)}]P$

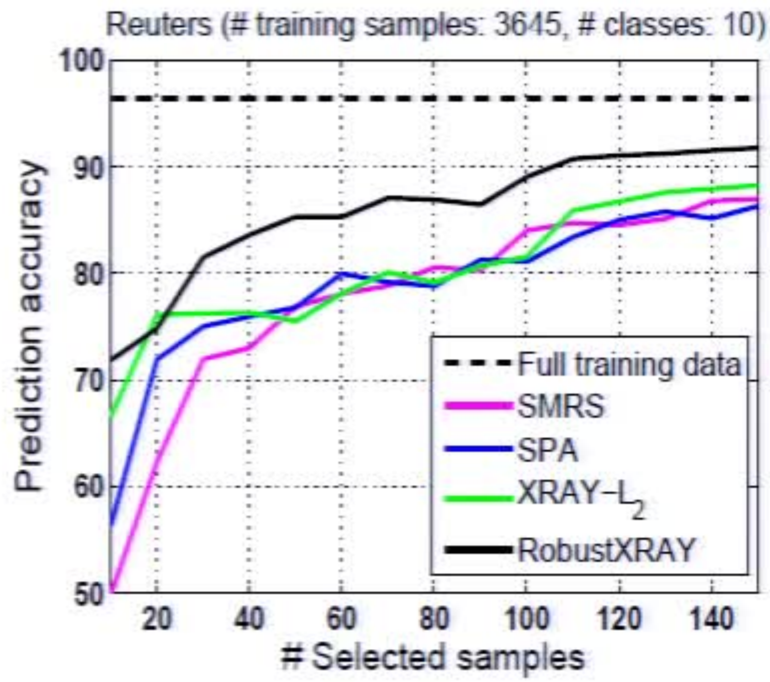




Exemplar Selection

- Proposed XRAY algorithms can also be used to select *exemplars* or *representatives*
  – video summarization, text corpus summarization
- Sparse Modeling Representative Selection (SMRS): Proposed in [Elhamifar et al, CVPR 2012]

$$\min_C \lambda \|C\|_{1,q} + \frac{1}{2}\|X - XC\|_F^2 \quad \text{s.t.} \quad 1^T C = 1$$



Summary

- Separability assumption makes the NMF problem tractable

  – reasonable assumption in applications like topic modeling, hyperspectral unmixing, etc.

- We develop a scalable family of algorithms for solving near-separable NMF problem under ℓ_1 and Bregman loss functions

  – incremental build-up of solution

  – outperforms competing approaches in performance and speed for video foreground-background separation and exemplar selection